GENOMICS

Division of Microbiology and Infectious Diseases

The Division of Microbiology and Infectious Diseases (DMID) supports a substantial program in genomics research, including sequencing of human pathogens and invertebrate vectors of diseases; applying genomic, functional genomics, and proteomic technologies to the study of microorganisms and infectious diseases; supporting genomic databases; and providing high-quality genomic reagents and technological resources to the scientific community.

Genome Sequencing

A genome is an organism's complete set of genes, encoded as a specific sequence of paired DNA bases. Recent advances in molecular biology have given researchers powerful methods that can quickly and accurately determine the complete DNA sequence of the whole genome of virtually any organism, including disease-causing microorganisms and the insect and other invertebrate vectors that can transmit them.

Whole-genome sequencing is an enormously powerful tool for understanding and defeating infectious diseases. For example, scientists can compare and contrast genomes to identify genes that are unique to a particular microbe. They can then target these genes with specific drugs, incorporate the products of these genes into experimental vaccines, and develop more sensitive diagnostic tests. Moreover, sequence information can reveal small genetic variations between different strains of a given pathogen. Researchers can use these subtle differences to determine which genes affect a pathogen's virulence, which genes are involved in the development of antibiotic resistance, and how a virulent or resistant strain spreads within a susceptible population; better understanding of these phenomena will help to improve disease diagnosis and patient care. Finally, understanding how microbial genes interact with one another and the human host during infection will lead to new strategies for drug therapies and vaccine development.

To capitalize on the tremendous potential of genome sequencing, NIAID has invested heavily in projects to sequence the genomes of medically important microbes. Sequencing technology has advanced to the point where NIAID working alone can fund the determination of a bacterial species; however, NIAID collaborates with other funding agencies to sequence the larger genomes of protozoan and fungal pathogens. To date, NIAID has completed the sequence of 92 genome-sequencing projects for 75 bacteria, 6 fungi, 9 parasitic protozoa, and 2 invertebrate vectors of infectious diseases. The bacterial species include those that cause anthrax, plague, tuberculosis, gonorrhea, chlamydia, cholera, strep throat, scarlet fever, and food-borne diseases. DNA sequencing projects have also been completed for the protozoan parasites Cryptosporidium parvum, *Entamoeba histolytica*, *Leishmania*, *Plasmodium falciparium*, *Giardia*, *Toxoplasma gondii*, *Trypanosoma cruzi*, and *Trypanosoma brucei* and the fungi *Aspergillus fumigatus*, *Aspergillus terreus*, *Cryptococcus neoformans*, and *Histoplasma capsulatum*. NIAID's data release policies ensure that both raw genome sequence data and associated assemblies and annotations are available to scientists around the world through deposition in an appropriate publicly searchable database (GenBank).

Study of the genomics of malaria has been particularly successful; for the first time, researchers have in hand the complete genetic sequences of the infectious organism, its natural host, and the insect that transmits it. In 2002, the International Malaria Genome Sequencing Consortium—funded in part by NIAID—published the genome sequence of *Plasmodium falciparum,* the parasite that causes the most severe form of malaria. NIAID also supported the rapid sequencing of the genome of *Anopheles*

gambiae, the mosquito that transmits the malaria parasite to humans. Researchers therefore now have the genome sequences of all three organisms involved in malaria—the mosquito vector, the malaria parasite, and the human host. This provides scientists with a unique opportunity to unravel the complex interactions between these three species on a molecular level. Indeed, NIAID-supported scientists already have taken advantage of this valuable genomic information to gain new insights into the molecular mechanisms involved in insecticide resistance, and to identify genes and gene products that are promising targets for new drug therapies.

The national biodefense effort has benefited substantially from genomic research as well, and NIAID has made a significant investment in sequencing microorganisms with the highest priority as potential agents of bioterrorism. For example, NIAID collaborated with the Office of Naval Research and the Department of Energy to sequence the genome of the Ames strain of *Bacillus anthracis,* the bacterium that causes anthrax. Other organisms sequenced include *Brucella suis, Burkholderia mallei,* two strains of *Clostridium perfringens, Coxiella burnetii,* and *Rickettsia typhi* with Defense Advanced Research Projects Agency funds; and six strains of *Bacillus anthracis*; *Bacillus cereus* strains; *Mycobacterium tuberculosis; Rickettsia rickettsii; Staphylococcus aureus; Yersinia pestis;* food-borne bacterial pathogens including diarrheagenic *E.coli, Vibrio cholerae, Shigella,* and *Salmonella;* and parasitic protozoa including *Cryptosporidium parvum (*human and bovine*), Giardia lamblia, Entamoeba histolytica,* and *Toxoplasma gondii.* In FY 2005, genome sequencing projects for *Burkholderia thailandensis*, five strains of *Burkholderia mallei* and *Burkholderia pseudomallei,* eight strains of *Escherichia coli*, two strains of *Yersinia pestis*, and one invertebrate vector of disease, *Aedes aegypti,* were completed. *Aedes aegypti* is the primary mosquito vector responsible for transmission of both yellow fever and dengue fever, and is the second invertebrate vector of infectious diseases sequenced by NIAID. Biodefense genome sequencing projects are currently being supported for additional strains of *Bacillus anthracis*, *Bacillus cereus*, *Burkholderia mallei, Burkholderia pseudomallei, Campylobacter, Coxiella burnetii, Escherichia coli, Entamoeba,* dengue viruses, *Francisella tularensis,* influenza, *Listeria, Mycobacterium tuberculosis, Ricinus communis, Shigella, Toxoplasma gondii, Vibrio cholerae, Yersinia pestis,* and the invertebrate vector, *Culex pipens*.

In FY 2005, NIAID continued to support the Influenza Genome Sequencing Project (*www.niaid.nih.gov/dmid/genomes/mscs/influenza.htm*), which is providing the scientific community with complete genome sequence data for thousands of human and animal influenza viruses. The influenza sequence data are being placed rapidly in the public domain through GenBank, an international searchable database, and NIAID's newly funded Bioinformatics Resource Center, with accompanying data analysis tools that enable scientists to study further how influenza viruses evolve, spread, and cause disease and might ultimately lead to improved methods of treatment and prevention. This newly generated sequence information is providing a larger and more representative sample of influenza virus genomes than was previously available to the public. This project has the capacity to sequence more than 200 genomes per month and is a collaborative effort among NIAID, the National Center for Biotechnology Information of the NIH's National Library of Medicine, The Institute for Genomic Research (TIGR), Wadsworth Center at the New York State Department of Health, Air Force Institute of Pathology, St. Jude Children's Research Hospital in Memphis, Centers for Disease Control and Prevention, Ohio State University, University of Maryland, Canterbury Health Laboratories (New Zealand), Los Alamos National Laboratories, and others. As of October 26, 2005, 463 complete genome sequences for influenza viruses had been released to GenBank, which include H1N1, H1N2, and H3N2 viral genomes collected from human clinical isolates.

Genomic Research

Obtaining the raw sequence of an organism's genome is only the first step in understanding it; annotating and organizing the sequence data are also required. Furthermore, the sequence data allow researchers to study an organism's proteome—the entire set of proteins that are encoded in the genome sequence. NIAID-supported investigators are applying such emerging genomic technologies to study microorganisms and infectious diseases. These studies include both basic research topics, such as the biology of a pathogen and the host's response to infection, and applied research such as development of medical diagnostics, drugs, and vaccines. Genomic technologies help scientists study infectious agents at the whole genome or proteome level. For example:

- Whole genome and proteome expression studies are being used to identify pathogen-specific genes and proteins involved in virulence, pathogenesis, and disease transmission.
- Proteomic technologies are being applied to both the pathogen and the host proteome to allow identification of candidate protein targets for new vaccines, therapeutics, and diagnostics.
- Genomic technologies are providing platforms for examination of genetic variation within and between species, strains, and clinical isolates, as well as for study of host responses to infection, vaccines, and antibiotic drugs.

Genomic Resources, Reagents, and Technologies

NIAID facilitates distribution of genomic resources and technologies to the research community for functional genomic analysis of microbial pathogens and supports the development of bioinformatics and computational tools that allow investigators to store and manipulate genomic and postgenomic data. In the past few years, NIAID has expanded its genomics activities and established comprehensive centers to provide the scientific community with needed reagents and resources to conduct basic and applied infectious diseases research. These centers include the NIAID Microbial Sequencing Centers, Pathogen Functional Genomics Resource Center (PFGRC), Bioinformatics Resource Centers, and Proteomics Research Centers.

NIAID continues to support the PFGRC at TIGR in Rockville, Maryland. PFGRC was established in 2001 to distribute to the research community a wide range of genomic and related resources and technologies for the functional analysis of microbial pathogens and invertebrate vectors of infectious diseases. Considerable progress has been made toward this goal, including the generation and distribution of 25 organism-specific DNA microarrays; the Center now includes microarrays for viruses, bacteria, fungi, and parasites. The available DNA microarrays include *Aspergillus fumigatus*, *Aspergillus nidulans, Chlamydia,* coronaviruses (animal and human), human SARS chip, *Helicobacter pylori, Mycobacterium smegmatis*, *Neisseria gonorrhoeae*, *Plasmodium falciparium*, *Pseudomonas aeruginosa, Staphylococcus aureus, Streptococcus agalactiae, Streptococcus pneumoniae, Trypanosoma brucei,* and *Trypanosoma cruzi.* In addition, organism-specific microarrays were produced and distributed for organisms considered agents of bioterrorism and include *Bacillus anthracis*, *Clostridium botulinum*, *Francisella tularensis, Giardia lamblia, Listeria monocytogenes, Mycobacterium tuberculosis, Rickettsia prowazekii, Salmonella typhimurium*, *Vibrio cholerae*, and *Yersinia pestis*. In addition, PFGRC has developed the methods and pipeline for generating organism-specific protein expression clones. Complete clone sets are now available for human SARS coronavirus, *Bacillus anthracis*, *Yersinia pestis*, and *Streptococcus pneumonia.* In addition, individual custom clone sets are available for more than 20 organisms upon request.

Further information is available at *www.niaid.nih.gov/dmid/genomes/pfgrc/default.htm.*

In FY 2003, NIAID awarded a contract to TIGR to support a Microbial Genome Sequencing Center to allow for rapid and cost-efficient production of high-quality microbial genome sequences; in early FY 2004, NIAID awarded a contract to the Massachusetts Institute of Technology to support a similar sequencing center. Genomes to be sequenced include microorganisms considered agents of bioterrorism (NIAID Category A, B, and C agents), microorganisms responsible for emerging and re-emerging infectious diseases, related pathogens, clinical isolates, and invertebrate vectors of infectious diseases. These sequencing centers have the capacity to respond to national needs and government priorities for genome sequencing, filling in sequence gaps and thus providing genome sequencing data for multiple uses, including forensic strain identification and identification of targets for drugs, vaccines, and diagnostics. In FY 2005, NIAID supported approximately 40 large-scale genome sequencing projects for additional strains of viruses, bacteria, fungi, parasites, viruses, and invertebrate vectors and included new projects for hepatitis C, coronaviruses, *Bacillus anthracis, Bacillus cereus, Bartonella bacilliformis, Burkholderia cenocepacia, Burkholderia dolosa, Campylobacter, Coxiella burnetii, Escherichia coli,* influenza, *Listeria, Pseudomonas aeruginosa, Shigella, Vibrio parahaemolyticus* , three strains of *Aspergillus,* additional strains of *Entamoeba, Plasmodium falciparium, Toxoplasma gondii,* additional sequencing of *Plasmodium vivax* and *Trichomonas vaginalis*, and one strain of *Ricinus communis*. Further information can be found at *www.niaid.nih.gov/dmid/genomes/mscs*.

The Malaria Research and Reference Reagent Resource Center *(www.malaria.mr4.org)* continued to provide expanded access to quality-controlled reagents for the international malaria research community in FY 2005.

Bioinformatics and Databases

In FY 2004, NIAID awarded eight contracts to establish Bioinformatics Resource Centers. These centers develop, populate, and maintain comprehensive relational databases to collect, store, display, annotate, query, and analyze genomic, structural, and related data for emerging and re-emerging pathogens, including those important for biodefense. The centers also develop and provide software tools to assist in data analysis. The databases these centers maintain are a valuable genomic resource, providing the scientific community with easy access to large amounts of genomic and related data and bioinformatics tools for data analysis. Further information is available at *www.niaid.nih.gov/dmid/genomes/brc/default.htm*.

Genomics and Proteomics

In the past several years, NIAID has awarded contracts for Biodefense Proteomics Research Centers, which develop and improve proteomic technologies and apply these technologies to pathogen and host cell proteomes for the discovery and identification of novel targets for the next generation of drugs, vaccines, diagnostics, and immunotherapeutics against microorganisms considered agents of bioterrorism. Eight centers have been funded to date; they focus on a range of NIAID category A, B, and C pathogens. Further information is available at *www.niaid.nih.gov/dmid/genomes/prc/default.htm.*

Division of Allergy, Immunology, and Transplantation

The Division of Allergy, Immunology, and Transplantation (DAIT) also supports genomics research. The human immune system is composed of complex networks of interacting cells, each programmed by precisely scripted genes. Underlying each immune response to a disease is a multistep pathway of interacting molecules influenced by an individual's unique genomic characteristics. The immune system plays a critical role in diseases such as rheumatoid arthritis; hay

fever; contact dermatitis; insulin-dependent or type 1 diabetes; systemic lupus erythematosus; and graft rejection of transplanted solid organs, tissues, and cells. Each of these diseases has an underlying genetic component.

Genomic research supported by DAIT is yielding insights into the functional and structural dimensions of immune system regulation, hypersensitivity, and inflammation in diseases such as asthma; the dysregulation of immune responses that results in autoimmune disease; and basic mechanisms of immune tolerance and graft rejection. This research is important in the following areas:

- **Asthma and allergic diseases.** DAIT-supported research on the genetics of asthma, hypersensitivity, inflammation, and T cell mediation increases understanding of the mechanisms underlying these immune responses, resulting in improved diagnostic, prevention, and treatment strategies. Through genomic research, DAIT-supported investigators discovered that interleukin-4 (IL-4), a cytokine produced by helper T cells and mast cells, stimulates antibody production by B cells in a series of reactions involving several genes. Further studies on IL-4 might provide a marker for measuring asthma risk and severity.

- **Autoimmune diseases.** DAIT supports research on type 1 diabetes and other autoimmune diseases that involve more than a single gene. Recent developments in genomics such as high-resolution DNA analysis and bioinformatics tools are making it possible to understand the underlying genetic causes of these complex diseases. For example, one approach compares the genes of individuals who have an autoimmune disease with those of healthy individuals to identify genetic and genomic differences that might be the underlying cause of disease. Between 10 and 20 distinct loci on the human genome might be responsible for susceptibility to type 1 diabetes. This knowledge will increase the ability to predict, diagnose, and treat this disease.

- **Transplantation.** DAIT-supported research on the genetics of graft rejection and immune tolerance is breaking new ground in the transplantation of solid organs, tissues, and cells for the prevention and treatment of disease. Genomic research funded by DAIT has identified surrogate markers of graft rejection in kidney transplant recipients. This research holds promise for the development of a noninvasive predictor of graft rejection based on gene expression analysis in urinary cells of transplant recipients.

- **Basic immunology research.** Basic research in immunology furthers understanding of the properties, interactions, and functions of the cells of the immune system and the genetic aspects of immune system regulation and provides information about essential structural immunobiology. Recent breakthroughs in the basic science of immunogenetics inform clinical immunology, which could lead to the development of new immune-based therapies. Examples of basic immunology research supported by DAIT include:

  - Use of large-scale gene- and protein-expression analysis tools to describe pathways of cellular activation;

  - Discovery of anti-inflammatory and immunosuppressive agents using DNA-based screening methods; and

  - Analysis of genomic databases of T cell receptors and immunoglobulin gene sequences to link structural, functional, and clinical information.

Multicenter Research Programs

DAIT supports several multicenter research programs that include significant genomic efforts aimed at understanding the underlying mechanisms of immune-mediated diseases.

Immune Tolerance Network (ITN). The ITN is an international consortium of more than 80 investigators in the United States, Canada, Europe, and Australia dedicated to the clinical evaluation of novel, tolerance-inducing therapies in autoimmune diseases; asthma and allergic diseases; and rejection of transplanted organs, tissues, and cells. The goal of these therapies is to re-educate the immune system to eliminate harmful immune responses while preserving protective immunity against infectious agents. To understand the underlying mechanisms of action of the candidate therapies and to monitor tolerance, the ITN has established state-of-the-art core laboratory facilities to conduct integrated mechanistic studies and to develop and evaluate markers and assays to measure the induction, maintenance, and loss of tolerance in humans. These core facilities include microarray analyses of gene expression, bioinformatics approaches to develop analytic tools for clinical and scientific datasets from the ITN-sponsored trials, enzyme-linked immunospot analyses of protein expression, and cellular assays for T cell reactivity. ITN is cosponsored by the National Institute of Diabetes and Digestive and Kidney Diseases and the Juvenile Diabetes Research Foundation International. More information on the ITN is available at *www.immunetolerance.org*.

Autoimmunity Centers of Excellence (ACEs). ACEs support collaborative basic and clinical research on autoimmune diseases, including single-site and multisite pilot clinical trials of promising immunomodulatory therapies. Clinical trials supported by ACEs include: a phase I/II clinical trial of anti-CD20 for treatment for lupus; phase I clinical trial of anti-tumor necrosis factor for treatment of lupus nephritis; preclinical study of DNase treatment, now underway with a follow-up phase Ib trial planned.

Multiple Autoimmune Disease Genetics Consortium (MADGC). MADGC is a repository of genetic and clinical data and specimens from families in which two or more individuals are affected by two or more distinct autoimmune diseases. This resource provides well-characterized material on 363 families to promote research aimed at discovering the human immune response genes involved in autoimmunity. More information can be found at *www.madgc.org*.

North American Rheumatoid Arthritis Consortium (NARAC). NARAC is a collaborative registry and repository of information on families with rheumatoid arthritis. The NARAC database contains information on 902 families, encompassing 1,522 patient visits. Of the 902 families, data for more than half have been validated, including 600 affected sibling pairs. The family registry and the repository samples should facilitate the characterization of the genes underlying susceptibility to rheumatoid arthritis and are available to all investigators. This registry is cosponsored by the National Institute of Arthritis and Musculoskeletal and Skin Diseases and the Arthritis Foundation. More information can be found at *www.naracdata.org*.

Primary Immunodeficiency Diseases Registry and Consortium. In FY 2003, the Primary Immunodeficiency Diseases Consortium was established with support from NIAID and the National Institute of Child Health and Human Development. The Consortium (1) provides leadership and mentoring; facilitates collaborations; enhances coordination of research efforts; and solicits, reviews, recommends, and makes awards for pilot or small research projects; (2) maintains a primary immunodeficiency diseases registry, which provides data to the research community about the clinical characteristics and prevalence of these diseases; and (3) is developing a repository of specimens from subjects with primary immunodeficiency diseases. Additional information on consortium activities is available at *www.usidnet.org*.